(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2003－36146
（P2003－36146A）
(43)公開日　平成15年２月７日(2003.2.7)

(51)Int.Cl.'		識別記号	FI		テーマコード'(参考)
Ｇ０６Ｆ	3/06	３０５	Ｇ０６Ｆ 3/06	３０５Ｃ	５Ｂ０１８
		５４０		５４０	５Ｂ０６５
	12/16	３２０	12/16	３２０Ｌ	

審査請求　未請求　請求項の数４　ＯＬ　（全６頁）

(21)出願番号　特願2001－221566(P2001－221566)

(22)出願日　平成13年７月23日(2001.7.23)

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(72)発明者　金子　誠司
神奈川県小田原市中里322番地２号　株式会社日立製作所ＲＡＩＤシステム事業部内
(74)代理人　100093492
弁理士　鈴木　市郎　（外１名）

Ｆターム(参考)　5B018 GA02 GA04 HA14 MA12
5B065 BA01 CA11 CA30 CC08 EA03
EA12 EA37 ZA15

(54)【発明の名称】　ディスクアレイ制御方式

(57)【要約】
【課題】　エラー検出能力の低いディスク制御部で、誤って旧データが読まれている障害を検出するとともに、付加チェックコードによってデータのエラーをも検出できるようにすること。
【解決手段】　複数のディスク装置を使って構成したディスクアレイ装置において、パリティグループを構成する複数のデータディスクに、各データディスクに書き込まれた各データに基づいて生成された各付加チェックコード２２０，２２１，又は２２２をそれぞれ書き込むとともに、パリティグループを構成するパリティディスクに、全ての各付加チェックコード２２０，２２１，及び２２２を書き込み、複数データディスクの１つと前記パリティディスクの付加チェックコードを読み出してエラーチェックを行うディスクアレイ制御方式。

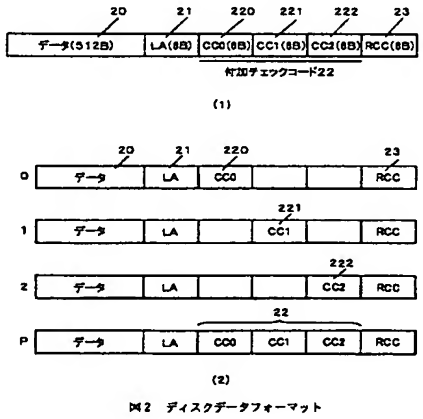

図2　ディスクデータフォーマット

【特許請求の範囲】
【請求項１】　複数のディスク装置を使って構成したディスクアレイ装置において、
パリティグループを構成する複数のデータディスクに、各データディスクに書き込まれた各データに基づいて生成された各付加チェックコードをそれぞれ書き込むとともに、
前記パリティグループを構成するパリティディスクに、全ての各付加チェックコードを書き込むことを特徴とするディスクアレイ制御方式。
【請求項２】　複数のディスク装置を使って構成したディスクアレイ装置において、
パリティグループを構成する複数のデータディスクに、各データディスクに書き込まれた各データに基づいて生成された各付加チェックコードをそれぞれ書き込むとともに、前記パリティグループを構成するパリティディスクに、全ての各付加チェックコードを書き込み、
前記複数データディスクの１つと前記パリティディスクの前記付加チェックコードを読み出してエラーチェックを行うことを特徴とするディスクアレイ制御方式。
【請求項３】　請求項２に記載のディスクアレイ制御方式において、
前記複数のデータディスクと前記パリティディスクに、書き込むべきデータのディスクアレイ内の位置情報をそれぞれ書き込み、誤った位置に書き込まれたデータのエラーチェックに前記位置情報を用いることを特徴とするディスクアレイ制御方式。
【請求項４】　請求項３に記載のディスクアレイ制御方式において、
パリティグループを構成するデータディスクとパリティディスクに、各データと各位置情報と各付加チェックコードとから生成されたチェックコードをディスクデータフォーマットの最後に書き込むことを特徴とするディスクアレイ制御方式。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、ディスクアレイ装置の障害検出方式に関し、特に、低価格で障害検出能力の低いディスク装置を組み合わせて、高信頼なディスクアレイ装置を構成する技術に関する。
【０００２】
【従来の技術】一般的に、ディスクは磁気記録媒体や光記録媒体を使用しているために、円板上の一部分が読み書きできなくなる障害が発生する。これをセクタ障害という。セクタ障害の原因としては、円板上の傷や磁性体の劣化等が考えられる。ディスク装置では、ＥＣＣ（Ｅｒｒｏｒ　Ｃｏｒｒｅｃｔｉｎｇ　Ｃｏｄｅ）を付加することで、ある程度のセクタ障害に対しては、復旧することが可能である。また、付加するＥＣＣ符号の種類により、セクタ中の何ビットまでのエラーを訂正することが可能であるかが変化する。
【０００３】低価格のドライブでは、上記のセクタ障害が起きる確率が比較的高いため、複数のディスクを並列に動作させることで高速制御を実現し、また、パリティと呼ぶ冗長データをパリティディスクと呼ぶ特定のディスクに格納することにより、万一、データを格納する１台のディスクが故障しても、他のディスクとパリティディスクのパリティとから故障したディスクのデータを再現することができ、耐ディスク障害信頼性を高めることができるＲＡＩＤと呼ばれるディスク制御の方法が提案された。
【０００４】このＲＡＩＤに関しては、「Ａ Ｃａｓｅ for Ｒｅｄｕｎｄａｎｔ Ａｒｒａｙｓ ｏｆ Ｉｎｅｘｐｅｎｓｉｖｅ Ｄｉｓｋｓ （ＲＡＩＤ）」：Ｉｎ Ｐｒｏｃ．ＡＣＭ ＳＩＧＭＯＤ，Ｊｕｎｅ １９８８（カリフォルニア大学バークレー校発行）に詳しい。ＲＡＩＤは、そのパリティの格納の方法によりレベル１からレベル５があるが、現在広く耐障害性を持つディスクアレイで用いられているのはレベル１とレベル５である。
【０００５】冗長構成を有する上記ＲＡＩＤ５構成を取るディスクアレイ装置では、セクタ障害が発生した場合、同一パリティグループに属する残りのディスク装置によってデータを復元し、交代領域への書き込みを行うことにより障害データを回復させていた。
【０００６】ここで、図３に示す４つのディスクへのデータ配置を例に取って、レイド５構成について述べると、ディスクアレイ制御装置に転送されてきたデータを０，１，２のように例えば５１２バイトの固定長に分けてディスク１，２，３に転送するとともに、パリティ部Ｐとしてデータ０，１，２のエクシクルーシブオアを取って５１２バイトの固定長を形成してディスク４に転送する。そして、このデータ０，１，２とパリティＰが同一パリティグループを形成しているのである。図３の矢印に示すように、データ０，１，２，パリティＰ，データ０，１，２，パリティＰ，…と続くディスク１～４へのデータ配置はレイド５の規格である。
【０００７】そして、ＲＡＩＤ５構成を取るひとつのディスクからデータを読み出そうとする場合、パリティグループを構成する全データディスクとパリティディスクからデータを読みとってエラーの検出をすることは可能であるが、複数のディスクからデータを読み出すため大きな性能低下を招く。
【０００８】このため、この様なＲＡＩＤ装置の場合、所定のデータにエラー検出用のコードを付け、その両方をデータとして書き込み、読み出しをおこなうことにより単体ディスクの読み出しでエラー検出が行える様に構成するのが通常である。即ち、データ０，１，２には、前述した５１２バイトの生データに加えて、ディスクのロジカルアドレスＬＡとチェックビットＣＨを付加して

いる。そして、このLAやCHによって単体ディスクでのエラー検出を行っている（図2を参照すると、図2のデータ20が生データに対応し、図2のLA21がロジカルアドレスに対応し、RCC23がチェックビットに対応するものである）。

【0009】また、特開平10-171608号公報には、データ部に対してアドレス情報と時間的要素を表す情報とを付加することが開示されている。このようなデータフォーマットを採用することで、磁気ディスク装置のキャッシュレジスタの古いデータを誤って読みだした場合でも磁気ディスク装置用データに変換する際にデータにタイムスタンプを追加記憶しておくので、データの読み出し又は書き込みの際にデータの異常を検出できるようになっている。

【0010】

【発明が解決しようとする課題】従来技術として説明した技術は、主にディスクの媒体に対する書き込み及び読み出し時のセクタ障害に代表されるエラーの検出、及びインターフェース関連のエラー検出とその回復を目的とする手段である。

【0011】一方、近年ディスクドライブはパーソナルコンピュータ用を中心に著しい価格の低下が進んだが、それにともない低価格化を重視してメディア（記録媒体）の障害に比して発生頻度の低い内部制御系のチェックの省略が行われたため、ディスクの制御部でのエラー検出能力は相対的に低くなっている。この場合でも、ディスク内部で発生したエラーが、例えばデータ系のビット化けなどの単純なデータ誤りとして書きこまれるならば、従来技術で説明したRAIDデータのエラー検出符号によって検出可能であるが、制御部用のデータを保持するディスク内のバッファに障害が発生した場合、書き込もうとしたデータを誤った位置に書き込む誤動作が発生し得る。

【0012】この種のエラーでは、書き込まれた先のデータ（誤った位置に書き込まれたデータ）も、上書きされないで残ったデータ（本来、上書きされて消去されるべき旧データ）も正しいチェックコードを持つ場合に、特に、書き込まれるべきデータが誤った場所に書き込まれたことによって書き込まれないで残った旧データは、これ自体単独では矛盾していないものなので、これが誤っていることを検出する手段が従来の技術では存在しない。

【0013】敷衍すると、512Bの生データ、位置情報のLA、チェックビットのCH、からなるデータ部と、生データから生成されたパリティデータ、LA及びCHからなるパリティ部と、で構成する従来のRAIDの場合、パリティ部（P）と全てのデータ部（0，1，2）との読み出しでエラーチェックすれば間違いなくエラー検出できるが、前述した読み出しの性能低下を考慮すると、データ部の読み出しでデータとチェックコードCHのチェックで問題なければ正のデータとしてホストに上げるという取り扱いをする場合がある。このような場合において、データが誤った位置に書き込まれた際に、書き込まれるべき位置に存するデータの読み出し指令に対して（ディスクアレイ制御装置は当該位置に新しいデータが書き込まれていると認識しているから）、上書きで消去されずに残っていた旧データが読み出されることとなる。旧データはそれ自体ではデータとCHとが一致するものであるから、この旧データの読み出しにエラー検出が掛からない（LAも問題がないのであるから）こととなる。

【0014】本発明の目的は、互いに共通するデータに基づいた付加チェックコードをデータ部とパリティ部の両方に保持して、データ部とパリティ部の2つのディスクでその付加チェックコードの検出を行うことによって、誤って旧データが読まれている障害を検出するとともに、付加チェックコードによってデータのエラーをも検出でき得ることにある。

【0015】

【課題を解決するための手段】本発明は、主として次のようなディスクアレイ制御方式を採用することによって、従来技術では為し得なかったエラー検出を実現する機能乃至作用を有するものである。

（1）パリティ部のパリティデータに加えて、パリティ付随情報として、RAIDを構成するデータ部に書き込んだ識別情報（図2に示すCC0、CC1又はCC2であり、データ20に基づいて生成されるもの）を付加チェックデータとして加えて書き込みを行う。

（2）常にデータ部とパリティ部の二つのディスク単位でデータの読み出しを行い、前記データ部の識別情報とパリティ部のパリティ付随情報とを比較し、誤って古いデータが読まれている障害を検出する。

【0016】本発明においては、読み出しには必ず二個のディスクにアクセスするため、読みだし性能は多少低下するが、古いデータを誤って読み出す障害を検出することができる。

【0017】以上のような機能乃至作用を果たすために、本発明は主として次のような構成を採用する。複数のディスク装置を使って構成したディスクアレイ装置において、パリティグループを構成する複数のデータディスクに、各データディスクに書き込まれた各データに基づいて生成された各付加チェックコードをそれぞれ書き込むとともに、前記パリティグループを構成するパリティディスクに、全ての各付加チェックコードを書き込み、前記複数データディスクの1つと前記パリティディスクの前記付加チェックコードを読み出してエラーチェックを行うディスクアレイ制御方式。

【0018】

【発明の実施の形態】本発明の実施形態に係るディスクアレイ制御方式について、図1、図2及び図3を用いて

以下詳細に説明する。図1は本発明の実施形態に係るディスクアレイ装置の概略的構成を示す図であり、図2は本実施形態におけるディスクに書き込むデータフォーマットを示す図であり、図3は本実施形態におけるディスク内のデータ配置を示す図である。

【００１９】図1において、１１０はホストとなるサーバ１０への通信路を示す。この通信路１１０には、一般にＳＣＳＩやファイバーチャネル等が使われる。図の１２は、本実施形態における制御を実施するディスクアレイ制御装置であり、ホスト通信制御部１２１、装置制御を行うプロセッサ部１２２、データのキャッシングを行うキャッシュ部１２３、ディスクの制御を行うディスク制御部１２４、から構成される。また、ディスクアレイ制御装置１２にはディスク１３が接続される。

【００２０】本実施形態では、図3を参照して、3個のデータに付きパリティデータを1つ用意し、4つのディスクでパリティグループを構成するＲＡＩＤ５構成を取るため、4の整数倍のディスク台数が必要であり、図3では簡単のため4台記載している。データとパリティ自体は、図3に示す様に分散されて配置されるため、データ用のディスクとパリティ用の専用ディスクに分かれているわけではない。

【００２１】図2は、ＲＡＩＤ装置に書き込むデータのフォーマットを示す。データ部とパリティ部は同じフォーマットを使用する。図2の２０はディスクの保持するデータ、図2の２１はそのデータのディスク中の位置情報、図2の２２は本発明の特徴である付加チェックコードデータ、図2の２３はデータ、位置情報、付加チェックコードから生成されるチェックコードである。

【００２２】本実施形態では、パリティディスクに書き込む付加チェックコードは、データ部の書き込みに用いる付加チェックコードと同一のもの（但し、データと位置情報から生成されたもの）を用い、またデータ部で用いる付加チェックコードは、パリティ部に用いる付加チェックコードを用いる。実際に書き込むデータに付随するチェックコードＲＣＣは、データ、位置情報、付加チェックコードから生成された全体のデータに対するチェックコードである。ここで、付加チェックコードは前述したようにデータ２０に基づいて生成されたチェックコードであることが特徴であり、ＣＲＣ（Ｃｙｃｌｉｃ Ｒｅｄｕｎｄａｎｃｙ Ｃｈｅｃｋ）コードを用いても良い。いずれもデータ２０に基づいたチェックコードであるから、この付加チェックコードを用いて、読み出しデータのエラーチェックも可能であり、更に、付加チェックコードの生成には既存のデータ及び／又は位置情報ＬＡを利用していて、特別な他の情報源を必要としたり、利用するものではな。また、付加チェックビットとしてデータを書き込んだ際に書き込みを一意に識別できる識別子を用いても良い。

【００２３】図3は、本実施形態のおけるディスク内のデータ配置を示す。パリティグループを構成する各ディスク３０，３１，３２，３３に、３つのデータ３４，３５，３６及びパリティ３７が図示の様に格納されるが、これは図示するように６４ＫＢｙｔｅｓ程度の細かい単位で別のディスクを使うようになっており、それによって負荷が高くなるパリティ部を全ディスクに分散させて負荷の均一化と性能向上を図っている。

【００２４】本実施形態の実際の動作を以下詳細に記載する。本実施形態では、ホスト１０側から通信路１１０経由でデータの書き込みが指定された場合、ホストから送られてきた書き込むべきデータを５１２Ｂｙｔｅｓ単位に分割して、まずキャッシュメモリ部１２３にバッファする。次に、制御プロセッサ１２２により、ディスク１３に書くべきデータを図2のフォーマットに従って生成する。ここで、データ２０は、ホストから送られてきたデータそのものであり、位置情報２１は、内部で管理している書き込むべきデータのディスクアレイ内の位置である。これは8バイトからなり、上位2バイトは、ディスクアレイ制御装置１２につながったディスクの通番（接続位置情報）であり、下位６Ｂｙｔｅｓは、ディスク内のデータの書き込み開始セクタ番号を示す。

【００２５】本実施形態では、書き込みの際には、対象とするデータが配置されるディスクと、パリティディスクの両方に対して同じ形式の図2に示す付加チェックコード２２を持つデータを書き込む。通常のＲＡＩＤの場合にもチェックコードとしてパリティデータを書き込むので、本実施形態が従来フォーマットと異なる点は、前記付加チェックコード２２を付随データとして書き込む点である。

【００２６】この時、書き込むデータの付加チェックコードデータ２２には、本実施形態のＲＡＩＤ構成に対応した、3つの付加チェックコードのフィールド２２０，２２１，２２２があるが、データ部の付加チェックコードは、自データの位置の付加チェックコードのみ更新、パリティデータの付加チェックコードは、書き込んだデータに対応する付加チェックコードのみ更新し、他の部分は元のままとする。

【００２７】例えば、パリティグループの第一のデータ領域３４にデータを書き込んだ場合（図3参照）、そのデータからパリティデータを再計算し、パリティデータの付加チェックコードをデータ領域３４のフィールド２２０の位置に書き、データから生成したチェックコードをパリティデータ３７のフィールド２２０の位置に書く。最後に、このデータ全体（データ２０、ＬＡ２１、付加チェックコード２２）に対して所定の検出能力を持ったチェックコード２３を生成し、データとして追加する。本実施形態では、８Ｂｙｔｅｓのチェックコードを付けており、５１２Ｂｙｔｅｓのデータにつき実際にディスクに書き込まれるのは５５２Ｂｙｔｅｓである。この５５２Ｂｙｔｅｓのデータをディスク制御装置１２か

ら実際のディスク１３に書き込む。

【００２８】以上説明したディスクデータフォーマットについて、図2の（２）を用いて再度説明する。ホストサーバ１０から送られてきたデータは、ディスクアレイ制御装置１２で、データ０，１，２のように固定長データに分割される。更に、パリティＰはデータ０，１，２からエクシクルーシブオアを取って作成された５１２Ｂのデータである。これらデータ０、１，２，パリティＰは、図3に示すようにディスク１～４に配置される。ここで、データ０のデータフォーマットについて、本実施形態の特徴である付加チェックコードＣＣ０が５１２Ｂのデータ、ＬＡ２１に基づいて作成されてＬＡ２１に続いて書き込まれ、更に、チェックコードＲＣＣ２３が５１２Ｂデータ、ＬＡ２１及びＣＣ０（２２０）に基づいて作成され書き込まれる。データ１及びデータ２についても、同様に付加チェックコードＣＣ１（２２１）及びＣＣ２（２２２）が図示のように書き込まれる。また、パリティＰについては、データ２０、ＬＡ２１に続いて、ＣＣ０（２２０）、ＣＣ１（２２１）、ＣＣ２（２２２）が全て書き込まれる。

【００２９】一方、本実施形態では読み出しの際の処理は通常と異なり、本実施形態に即して必ず対象とするデータが配置されるディスク３０～３３の一つとパリティデータ３７の両方を読み込む。読み出した付加チェックコード２２が読み出したデータの対応する位置の付加チェックコードと一致しているかを確認してエラーの検出を行う。本実施形態の場合、エラー検出は以下の判定で行う。

【００３０】（１）チェックコード２３とデータ２０が不一致の場合には、そのデータはエラーである。これはディスク内部の書き込み時のデータ系の障害によって正しいデータが書き込まれなかったものであると考えられる。

【００３１】（２）チェックコード２３とデータ２０は一致しているが、読み出されたデータに埋め込まれた位置情報２１と、そのデータ２０が本来あるべき位置（ディスクアレイ制御装置が書き込みを指令した位置）が一致していなかった場合は、その不一致となっているデータはエラーである。これはディスク１３の内蔵制御部障害によって誤った位置に書き込まれた先のものが読み出されたケースである。

【００３２】（３）付加チェックコード２２が一致していなかった場合には、まずパリティグループのデータを読み、そのデータから不一致の各データが誤っていたことを仮定した二通りのデータ再構築を行い、その結果と付加チェックコード２２を比較して矛盾している側が誤ったデータである。これは、制御部障害によって書き込まれるべきデータが誤った位置に書き込まれたため、旧データが読み出されているケースである。即ち、図2の（２）の場合を例にすると、データ０とパリティＰを読み出して付加チェックコードが不一致の場合、データ０とパリティＰのいずれかがエラーである。そこで、データ１とパリティＰ、データ２とパリティＰ、の組み合わせで付加チェックコードの一致をみて、データ０とパリティＰのいずれかのエラーを検出する。データ０がエラーであると解ると、データ１、データ２及びパリティＰからデータ０の回復を行う。

【００３３】（４）不一致やチェックコードとの不整合がなかった時にはデータは正しく、そのまま用いて良い。

【００３４】上記判定によりデータのエラーが検出された場合には、改めてアレイを構成する全データを読み込み、そのデータから障害によって失われたデータを回復する。リカバリ時は、付加チェックコードデータを読み出したデータから再構築することを除けば通常のＲＡＩＤデータ回復手順と全く同じ処理である。

【００３５】本実施形態では、データディスク３つに対してパリティディスク１つを置く構成を取ったが、チェックコード２３のフィールド数を増減させることにより他のＲＡＩＤ構成に対応させることも容易に可能である。

【００３６】本実施形態では、識別子としてＣＲＣ等の情報を用いたが、これはＲＡＩＤ装置としてのエラー回復処理が容易となるためである。例えば、データ部を書き込んだ書き込み通算番号をパリティ部に保持する様にしてもＲＡＩＤ５では、同様のエラー検出能力を得ることができる。この場合３台以下のパリティグループ構成では、どちらのデータが正しいかを知ることができないという難点があり、チェックコードを更に追加する必要がある。

【００３７】

【発明の効果】本発明によれば、ディスク制御部でのエラー検出能力の低いディスクを使いながら、エラー検出能力を備え、かつ性能低下を抑えたＲＡＩＤ装置を構成することができる。

【図面の簡単な説明】

【図１】本発明の実施形態に係るディスクアレイ装置の概略的構成を示す図である。

【図２】本実施形態におけるディスクに書き込むデータフォーマットを示す図である。

【図３】本実施形態におけるディスク内のデータ配置を示す図である。

【符号の説明】

１２　ディスクアレイ制御装置
１３　ディスク
２０　データ
２１　位置情報
３４，３５，３６　データ領域
３７　パリティデータ領域
１１０　通信路

121 ホスト通信制御部
122 制御プロセッサ
123 キャッシュメモリ部
124 ディスク制御部

【図1】

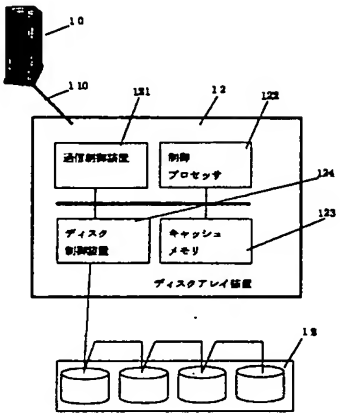

図1 ディスクアレイ装置

【図2】

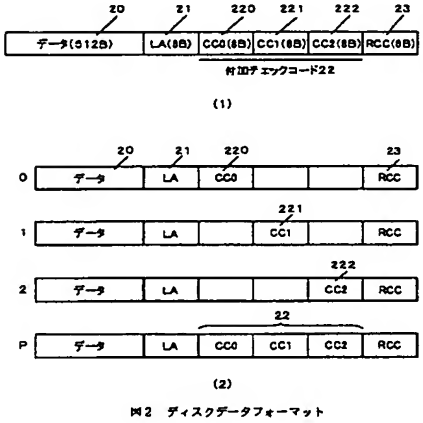

図2 ディスクデータフォーマット

【図3】

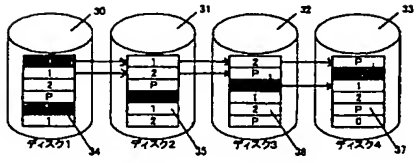

図3 ディスクデータ配置